

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2000-134214

(43)Date of publication of application : 12.05.2000

(51)Int.Cl. H04L 12/28
H04L 12/46
H04L 12/56

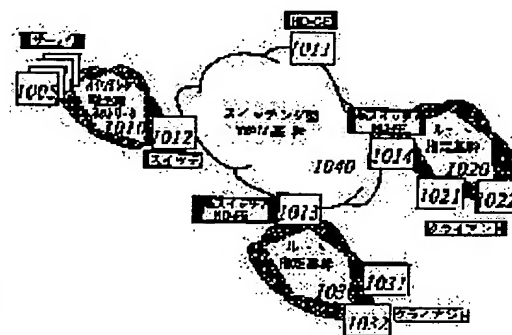
(21)Application number : 10-301470 (71)Applicant : INTERNATL BUSINESS MACH
CORP <IBM>
(22)Date of filing : 22.10.1998 (72)Inventor : GILLENE CRAND
HUNT GUERNEY DOUGLAS
HOLLOWAY
ERIC MICHEL LEVY-ABENOORU
DANIEL GEORGES JEAN-MARIE
MODYU

(54) DISTRIBUTION TYPE SCALABLE SYSTEM SELECTING SERVER FROM SERVER CLUSTER TO SELECT SWITCHING PATH LEADING TO THE SELECTED SERVER

(57)Abstract:

PROBLEM TO BE SOLVED: To build up a scalable CP/IP service from a switch or a network by transmitting data relating to a request of a client by a transfer engine to a selected server through a switching type connection not through a control engine.

SOLUTION: Network dispatcher transfer engines (FE) 1013, 1014 receive a request of a client and the FE 1013, 1014 transmit a request to a network dispatcher control engine(CE) 1011 in response to the request of the client to select a server and a corresponding switching address from a cluster. The CE 1011 selects the server to send a corresponding switching address to the FE 1013, 1014, and the FE 1013, 1014 transmits data relating to the request of the client to the selected server through a switching type connection relating to the switching address. The switching type connection is not required to pass through the CE 1011.



LEGAL STATUS

[Date of request for examination] 27.07.1999

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number] 3327850

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開2000-134214

(P2000-134214A)

(43)公開日 平成12年5月12日(2000.5.12)

(51)Int.Cl. ⁷	識別記号	F I	テマコード*(参考)
H 0 4 L	12/28	H 0 4 L 11/20	G 5 K 0 3 0
	12/46	11/00	3 1 0 C 5 K 0 3 3
	12/56	11/20	1 0 2 D

審査請求 有 請求項の数22 O L (全 19 頁)

(21)出願番号 特願平10-301470

(22)出願日 平成10年10月22日(1998.10.22)

(71)出願人 390009531

インターナショナル・ビジネス・マシーンズ・コーポレーション

INTERNATIONAL BUSINESS MACHINES CORPORATION

アメリカ合衆国10504、ニューヨーク州
アーモンク (番地なし)

(72)発明者 ジレーヌ・克蘭ド

フランス国06100、ニース、レ・カレブ、
アヴニュ・アンリ・デュナン 70

(74)代理人 100086243

弁理士 坂口 博 (外2名)

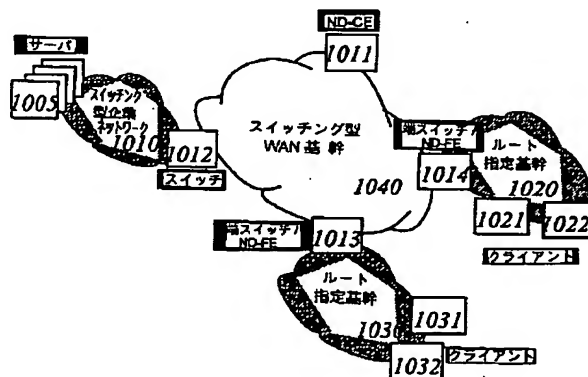
最終頁に続く

(54)【発明の名称】 サーバ・クラスタからサーバを選択し、選択されたサーバへのスイッチング経路を選択するための分散型スケーラブル装置

(57)【要約】 (修正有)

【課題】 ATMスイッチまたはATMスイッチング型のネットワークなどのスイッチまたはスイッチング型のネットワークの中からクラスタ・ウェブ・サーバなどのスケール可能なTCP/IPサービスを構築する。

【解決手段】 分散化された大型TCPルータはATMスイッチング型のネットワークを用いて構築されうる。スケールされたサービスは単一のサービスとしてクライアントに与えられる。クライアントはスイッチまたはスイッチング型のネットワークに直接または間接に接続される。1バージョンは制御エンジン(CE)および転送エンジン(FE)と言う2つの要素を含む。CEはサーバに接続を割り当て、割り当てられたサーバに関する情報を転送し、またFEへの接続を割り当てる機能を持つ。FEはCEから受け取った割り当てを適用してTCP接続をスイッチ型ATM接続にマップする。終了時に、FEは接続終了事象をCEに戻す。



【特許請求の範囲】

【請求項1】スイッチング型のネットワークを含むクライアント-サーバ・システムにおいてサーバのクラスタからあるサーバを選択すると共に選択されたサーバに至るスイッチされる経路を選択するための方法であって、転送エンジンによりクライアントの要求を受け取るステップと、クライアントの要求に回答して、クラスタからサーバを選択すると共に対応するスイッチング・アドレスを選択するため、転送エンジンにより要求を制御エンジンに送るルートを指定するステップと、制御エンジンによりサーバを選択して対応するスイッチング・アドレスを転送エンジンに送るステップと、転送エンジンにより、クライアントの要求に関連されたデータを、スイッチング・アドレスに関連するスイッチング型接続を通して、制御エンジンを通ることなく選択されたサーバに送るステップと、を含む方法。

【請求項2】スイッチング・アドレスを転送エンジンに送る前記ステップは、制御エンジンによりサーバ選択基準および転送エンジンがこの基準を使用できる条件を転送エンジンに送り、転送エンジンにより受け取られる後続のクライアントの要求に対して、転送エンジンによりこの要求に現存するスイッチング型接続が関連しているか否かを調べ、現存する接続が存在するならば、転送エンジンにより現存する接続を通してこの要求を送り、現存する接続が存在しないならば、転送エンジンにより前記基準に基づいて宛先をローカルに選択する、ステップを含む、請求項1に記載の方法。

【請求項3】転送エンジンによりスイッチング・アドレスに基づいて、スイッチング型ATM接続にTCP接続をマップするステップを更に含む請求項1に記載の方法。

【請求項4】制御エンジンから遠隔の場所に複数の転送エンジンを配置してこの転送エンジンをスイッチング型ネットワークの端に分散させ、前記転送エンジンの各々により制御エンジンの指示の下でTCP接続を分散させるステップ、を更に含む請求項3に記載の方法。

【請求項5】分散された複数の転送エンジンをスイッチ組織を介して制御エンジンに接続させ、制御エンジンの指示の下で、分散された各転送エンジンによりスイッチング型接続をマップするステップ、を更に含む請求項3に記載の方法。

【請求項6】クライアント-サーバ・システムが制御エンジンに障害が生じたときのため一次制御エンジンおよびバックアップ制御エンジンを含んでおり、

前記方法が、一次制御エンジンの障害を検出するステップと、前記検出するステップに回答して、バックアップ制御エンジンが一次制御エンジンを引継ぎ、それが新たな一次制御エンジンであることを転送エンジンに通知するステップと、を含む請求項1に記載の方法。

【請求項7】クライアント-サーバ・システムが複数の転送エンジンを含んでおり、前記方法が、

クライアントの要求に回答して、構成情報を用いて、一次転送エンジンが障害を起こしたとき選択される1つ以上のバックアップ転送エンジンを構成に組み入れるステップと、

一次転送エンジンが障害を起こしたとき、活動中のクライアント接続を中断することなくバックアップ転送エンジンにデータをルート指定するステップと、を更に含む請求項1に記載の方法。

【請求項8】障害を起こした転送エンジンが回復したことを判定し、回復した転送エンジンを更新するステップと、

新たな要求が回復した転送エンジンにルート指定され、現存する接続のためのパケットをクライアントに中断を与えることなく一次転送エンジンとしての回復した転送エンジンに再ルート指定するようにネットワークを更新するステップと、を更に含む請求項7に記載の方法。

【請求項9】クライアントがスイッチング型のネットワークに直接取り付けられ、クライアントが転送エンジンを含んでいる請求項1に記載の方法。

【請求項10】クライアントがインタネットを介してスイッチング型のネットワークに取り付けられる請求項1に記載の方法。

【請求項11】スイッチング型のネットワークを含むクライアント-サーバ・システムにおいてサーバのクラスタからあるサーバを選択すると共に選択されたサーバに至るスイッチング経路を選択するためのシステムであって、

クライアントの要求を受け取って要求を制御エンジンにルート指定し、クライアントの要求に回答してクラスタからサーバおよび対応するスイッチング・アドレスを選択するための手段と、

前記制御エンジンは、サーバ選択要求に回答してサーバを選択するための制御エンジン手段、および対応するスイッチング・アドレスを転送エンジンに送るための制御エンジン手段を含むことと、

前記転送エンジンは前記スイッチング・アドレスに関連するスイッチング接続を通して制御エンジンを通ることなくクライアントの要求に関連するデータを選択されたサーバに送るための転送エンジン手段と、

を含むシステム。

【請求項12】前記スイッチング・アドレスを転送エンジンに送るための制御エンジン手段は、サーバ選択基準および転送エンジンが該基準を使用できる条件を転送エンジンに送るための制御エンジン手段を含み、その後受け取られるクライアントの要求に対して、転送エンジンが、

この要求に関連する現存するスイッチング接続が存在するか否かを判定するための転送エンジン手段と、

現存する接続が存在する場合、その接続を通して要求を送るための転送エンジン手段と、

現存する接続がない場合、前記基準に基づいて宛先サーバをローカルに選択するための転送エンジン手段と、を更に含む請求項11に記載のシステム。

【請求項13】スイッチング・アドレスに基づいてTCP接続をスイッチング型ATM接続にマップするための転送エンジン手段を更に含む請求項11に記載のシステム。

【請求項14】制御エンジンから遠隔の場所にあるスイッチング型ネットワークの端に分散された複数の転送エンジンを含み、

前記転送エンジンの各々が制御エンジンの指示の下でTCP接続を分散させることを特徴とする請求項13に記載のシステム。

【請求項15】分散された複数の転送エンジンをスイッチ組織を介して制御エンジンに接続させ、制御エンジンの指示の下で、分散された各転送エンジンによりスイッチング型接続をマップする、ことを特徴とする請求項13に記載のシステム。

【請求項16】制御エンジンに障害が生じたときのために一次制御エンジンおよびバックアップ制御エンジンを含んでおり、

一次制御エンジンの障害を検出する手段と、

前記検出する手段に応答して、バックアップ制御エンジンが一次制御エンジンを引継ぎ、それが新たな一次制御エンジンであることを転送エンジンに通知するためのバックアップ制御エンジン手段と、

を更に含む請求項11に記載のシステム。

【請求項17】複数の転送エンジンを含んでおり、クライアントの要求に回答して、ネットワークで得られる構成情報を用いて、一次転送エンジンが障害を起こしたとき選択されうる1つ以上のバックアップ転送エンジンを構成に組み入れるための手段と、

一次転送エンジンが障害を起こしたとき、活動中のクライアント接続を中断することなくバックアップ転送エンジンにデータをルート指定するための手段と、

を更に含む請求項11に記載のシステム。

【請求項18】障害を起こした転送エンジンが回復したことを判定し、回復した転送エンジンを更新するための

手段と、

新たな要求が回復した転送エンジンにルート指定され、現存する接続のためのパケットをクライアントに中断を与えることなく一次転送エンジンとしての回復した転送エンジンに再ルート指定するようにネットワークを更新するための手段と、

を更に含む請求項17に記載のシステム。

【請求項19】クライアントがスイッチング型のネットワークに直接取り付けられ、クライアントが転送エンジンを含んでいる請求項11に記載のシステム。

【請求項20】クライアントがインタネットを介してスイッチング型のネットワークに取り付けられる請求項11に記載のシステム。

【請求項21】サーバがルート指定されるネットワークに取り付けられる請求項11に記載のシステム。

【請求項22】転送エンジン、制御エンジン、およびスイッチ組織が単一の装置に共存する請求項15に記載のシステム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】 インタネット・エンジニアリング・タスク・フォース (IETF) のNBMA (non-broadcast multiple access) でのインタネットワーキング (Internetworking over NBMA-ION) のワーキング・グループはインタネット・プロトコル (IP) スイッチングに対する3つの異なる提案を現在検討中である。これらのアーキテクチャは2つの方法により要約することができる。即ち先ず、イプシロンスイッチング方法は非同期転送モード (ATM) 接続をインタネット・プロトコル・フローに関連づけ、次のものは接続を出口ルータのルートに関連づける。

【0002】

【従来の技術】 ATMはこの分野で良く知られている。概観すると、ATMは歴史的にはB-ISDN (Broadband and Integrated Services Digital Network) の開発に基づいている。ATMはB-ISDNの伝送モードとして選ばれたパケットを多重化しスイッチングする方法である。ATM、即ち高速デジタル伝送、はパケット・スイッチング技術を用いるが、「非同期」伝送とは関係がない。(例えば、PDH, Broadband ISDN, ATM and All That: A Guide to Modern WAN Networking, and How It Evolved, Paul Reilly著, Silicon Graphics Inc. April 4, 1994発行を参照されたい)。ATMパケットはセルと呼ばれ、各セルは5バイトのヘッダおよび48バイトのデータを有する。ATMパケット・スイッチングは、ATMパケットが仮想的経路および仮想的回路と呼ばれる予め設定されたルートを辿るという点において従来のパケット・スイッチングと異なる。ATMはいかなる特定の物理的伝送媒体にも依存しないが、伝送媒体が主として光ファイバであるときにはエラー率および損失率が

非常に少なく、従って再伝送は行われない。"Asynchronous Transfer Mode Tutorial", Northern Telecom, <http://www.webproforum.com/nortel2/index.html>, (6/10/98)を参照されたい。

【0003】トランスミッション・コントロール・プロトコル／インタネット・プロトコル(TCP/IP)およびATMにおけるTCP/IPの使用はこの分野で良く知られている。D.E. Comer著、"Internetworking with TCP/IP: Principles, Protocols, and Architecture", Englewood Cliffs, N.J., Prentice Hall発行(1988)を参照されたい。トランスミッション・コントロール・プロトコル(TCP)スイッチングは異なるATMルータの間で接続を割り当てることによって、例えば予め定義された仮想経路インディケータ／仮想チャネル・インディケータ(VPI/VCI)を用いて働くが、この方法は接続当たり交換される所与の量のパケットが効率的であることを必要とする。取りうる別の処理方法はイブシロンIPスイッチング方法を用いることである。

【0004】ワールドワイド・ウェブ上のトラフィックは特に評判の良い(ホット)所では指数的に増大している。従ってスケーラブルなウェブ・サーバを提供することが重要である(例えば、Goldschmidt, G. および Hunt, G. 著、"Net Dispatcher a TCP Connection Router" IBM Research Report, 1997, および Dias, D.M., Kish, W., Mukherjee, R., Tewari, R. 著、"A Scalable and Highly Available Web Server", Proc. 41st IEEE Computer Society Intl. Conf. (COMPCON), 1996, Technologies for the Information Superhighway, pp85-92, Feb. 1996)を参照されたい。スケーラブル・ウェブ・サーバにおいて負荷バランスを与える公知の方法はネットワーク・ディスパッチャを用いることである(例えば、米国特許第5,371,852号、およびAttanasio, Clement R., Smith, Stephen E. 著、"A Virtual Multi-Processor Implemented by an Encapsulated Cluster of Loosely Coupled Computers", IBM Research Report RC 18442 (1992)を参照されたい)。ここではネットワーク・ディスパッチャ(ND)のアドレスだけがクライアントに与えられ、ネットワーク・ディスパッチャがクラスタ(いわゆる仮想的カプセル化クラスターVEC)内のノードの間で、ラウンド・ロビン式に、またはノードの負荷に基づいて、入来要求を分配する。継続中の米国特許出願番号08/861,749には一般化されたネットワーク・ディスパッチャの例が開示されており、これはネットワーク相互間全体の任意の場所に所在するノードへのルート決定を可能にする。

【0005】インタネットの基幹ネットワークはスイッチング型ATMインフラストラクチャに移行しつつある。同時に、非常に大きなサーバ(メイン・フレーム、メイン・フレーム・クラスタ、またはその他のタイプのクラスタの如何を問わず)が、帯域幅および要求時スル

ープットの画期的な成長を扱うためATMリンクを介してインタネット基幹に接続されようとしている。

【0006】この意味において、IETFはATMスイッチの単純、高速、効率的な処理能力の利点を活用するための種々の代替案を検討中である。種々の代替案に共通するものは、IPヘッダに基づくルート決定をATMヘッダに基づくスイッチング決定と置き換えてすべての中間のホップ(クライアントおよびサーバ以外のすべてのホップ)処理を単純化するための動的な方式である。このことは究極的にはエンドポイント(つまり、クライアントおよびサーバ)だけがIPパケット(IPレイヤー、TCPレイヤーなど)を処理し、エンドポイント間の経路上のその他のすべてのホップがATMパケットをスイッチすることを意味する。代替案の或るものはいわゆる「ショートカット」方法も検討しており、これは物理的接続性が許容するときには中間のホップの幾つかをバイパスする機構である。インタネット業界で検討されている解決策には、ネクスト・ホップ・レゾリューション・プロトコル(NHRP)、イブシロンIPスイッチング・プロトコル(IFMPおよびGSMP)、タグ・スイッチング、およびIBMのアグリゲート・ルート・ベース・IPスイッチ(ARIS)がある。NHRPについては、"Next Hop Resolution Protocol (NHRP)", The Internet Society, Network Working Group, RFC 2332 (1998)を参照されたい。

【0007】全面的にまたは部分的であってもスイッチング型のネットワークにおいては、サーバのクラスタに対し従来型のフロント・エンド(ネットワーク・ディスパッチャなど)を稼働させるホップはIETFで検討されている革新的なアプローチ全体と相容れないものとなる。IPおよびTCP分野を調べることが必要であり、その他のどのようなホップもルート決定を行うのにIPを考慮するのを避けようとすることになるであろう。

【0008】

【発明が解決しようとする課題】上述の必要性に従って、本発明はサーバのクラスタに対するフロント・エンドにスイッチング能力を与えて、パケットがサーバまで、また、サーバに戻るように、またはクライアントに最も近いスイッチへと、スイッチされるようにする特徴を有する。

【0009】

【課題を解決するための手段】本発明の1態様は1つの要素、すなわち、制御エンジン(CE)および転送エンジン(TE)を含む。制御エンジンはサーバへの接続を割り当てて、割り当てられたサーバに関する情報を伝送し、そして転送エンジンに接続する機能を有する。各転送エンジンは制御エンジンから受け取った割り当てを適用してスイッチング型ATM接続へのTCP接続をマップする。転送エンジンは終了時に接続終了事象を制御エンジンに戻す。

【0010】スイッチング型のネットワークを含むクライアントサーバ・システムにおいてサーバのクラスタからサーバを選択し、選択されたサーバに至るスイッチされた経路を選択する方法の例は以下のステップを含む。即ち、転送エンジン（FE）がクライアントの要求を受け取るステップと、クライアントの要求にตอบสนองしてFEが制御エンジン（CE）に要求を送ってクラスタからサーバおよび対応するスイッチング・アドレスを選択するステップと、CEがサーバを選択して対応するスイッチング・アドレスをFEに送り、FEがクライアントの要求に関連するデータをスイッチング・アドレスに関連するスイッチング型接続を通して選択されたサーバに送るステップとが含まれる。ここでスイッチング型接続はCEを通る必要はない。

【0011】CEがスイッチング・アドレスをFEに送るステップは更に以下のステップを含む。即ち、CEがサーバ選択基準およびFEがその基準を使用することができる条件をFEに送るステップと、FEにより受け取られたその後のクライアント要求に対して、現存するスイッチされた接続がこの要求に関連しているかどうかをFEが判定するステップと、現存する接続があるならば、現存するスイッチされた接続を通してFEがこの要求を送るステップと、現存するスイッチされた接続がないならばFEが基準に基づいて宛先をローカルに選択するステップと、が含まれる。

【0012】本発明はその他の特徴を有しており、これは中央に集中されたクラスタ・フロント・エンド（クラスタ・サーバまたはディスパッチャとも呼ばれる）を通るパケットの経路を最小にし、ルート決定機能の幾分かをスイッチング型のネットワークの端に分配することによって隘路が生じる可能性を都合良く低減する。また、この方法はディスパッチャに2ホップ以上離れたサーバを管理させ、また伝送処理を分配することによってシステム全体の頑強さおよび性能を増大させる。例えば、FEはスイッチング・アドレスに基づいてTCP接続をスイッチング型ATM接続にマップすることができる。CEから遠くに離れた複数のFEを設けてFEがスイッチング型のネットワークの端に分布され、各FEがCEの指図に従ってTCP接続を分配するようにすることもできる。別の例では、複数の分散されたFEがスイッチ組織を介してCEに接続されることも可能であり、分散された各FEはCEの指図に従ってスイッチング型接続をマップする。

【0013】本発明の1実施例は現存するTCP接続ルータのすべての能力を含んでいる。これには利用可能性が高いことおよびフォールト・トレランスが含まれる。例えば、1997年9月15日出願、継続中の米国特許出願（出願人整理番号Y0997-232）を参照されたい。別の実施例は現存するTCP接続型のルータの、サーバ・クラスタに加わる負荷をバランスさせるためのフィード

バック機能の幾つかを含んでいる。

【0014】本発明によるフォールト・トレランスの特徴を含む方法の1例は一次CEと、この一次CEが故障したときのためのバックアップCEとを含んでいる。この方法は、一次CEの障害を検出するステップと、障害の検出にตอบสนองしてバックアップCEが一次CEを引き継いでFEにそれが新しい一次CEであることを通知するステップとを含む。

【0015】複数のFEを含む別の例はクライアントの要求にตอบสนองしてネットワークで入手できるコンフィギュレーション情報を用いて一次FEが障害を起こしたときに選択されるバックアップFEを1つ以上構成に組み入れるステップと、一次FEが障害を起こしたとき活動中のクライアント接続を中断することなくバックアップFEにデータを送るステップとを含む。追加のステップとして、障害を起こしたFEが回復したことを判定するステップと、回復したFEを現状に更新するステップと、ネットワークを更新して新たな要求が回復したFEに送られるようにするステップと、クライアントを中断することなく現存する接続のパケットを一次FEとして回復したFEに経路変更するステップとが含まれても良い。

【0016】本発明は取り付けられた（直接または間接に）サーバによって提供されるサービスをスケールする汎用装置を開発するためのシステムおよび方法を含む。この汎用装置は強化されたスケール能力を与えるためのスイッチを利用する。この装置の1例において、分散大型TCPルータがATMスイッチング型のネットワークを用いて構築される。スケールされたサービスはクライアントに対する単一のサービスとして提供されうる。これらのサービスはスイッチ組織に直接または遠隔地から取り付けられ得る。

【0017】幾つかの利点として、次のものが挙げられる。

- ・ インタネット・サービスをスケールする現在のいかなる手法と比べても最高の容量およびスループットを有する。
- ・ この手法はフォールト・トレランスおよび高い利用可能性を有する。
- ・ この手法はサーバおよびクラスタ・サーバが同一サブネットに共存することを強制すると言のような制約を持たない。サーバはATM組織に直接取り付けられまたはルート経由のネットワークを介して取り付けられ得る。
- ・ スwitch組織はネットワーク内にあって良く、または高度にスケール可能な並列コンピュータであっても良く、或いはその他任意のアプリケーション（電話など）であって良い。

【0018】

【発明の実施の形態】図1はスイッチング技術を含むスイッチング型のネットワークに配置された本発明の例を示す。一般に、スイッチング・ワイド・エリア・ネット

トワークの基幹1040は、フレーム・リレー、ATM、またはX.25を含む（これらに限られるものではないが）任意のスイッチ技術により具体化された従来型のWAN（Wide Area Network）である。同様にスイッチされる企業ネットワーク1010は、フレーム・リレー、ATM、またはX.25等の任意のスイッチ技術により具体化された従来型の企業ネットワークである。従来通り、基幹ネットワークはインタネットに対する中央の相互接続を作る任意のネットワークである。全国的基幹は通常はWANであり、企業基幹は通常はLAN（Local Area Network）またはLANの組合せである。

【0019】本発明は周知のネットワーク・ディスパッチャ（ND）を含むことが望ましい。これはTCP接続のソフトウェア・ルータであり、複数のTCPサーバ間で負荷のバランスをも支援する。しかしながら、任意のTCP接続ルータや負荷バランス論理が本発明に採用できることは当業者にとって明らかであろう。図に示されたように、本発明は2つの主体であるネットワーク・ディスパッチャ制御エンジン（ND-CE）1011およびネットワーク・ディスパッチャ転送エンジン（ND-FE）1013、1014を含む。ND-CE1011はサーバ1005への接続を割り当てて、そのサーバに関する情報を送り、またND-FE1013、1014への接続を割り当てる。ND-FE1013、1014の各々はND-CE1011から受け取った割り当てを適用してスイッチング（例えばATM）接続へのTCP接続をマップする。ND-FEは最終的には接続終了事象をND-CE1011に戻す。この方法がND-CE1011とND-FE1013、1014との間で情報を伝播させるのに用いられるプロトコルに無関係であると共に、これらの機能の物理的所在にも無関係であることは当業者に明らかであろう。

【0020】転送エンジン（FE）は次の2つの極端に至るまでネットワークのどこにあって良い。

1. クライアント側：TCP接続がクライアントからサーバへ全体的にスイッチされるようにする。この手法は最も効率的であり、クライアントがスイッチング型のネットワーク（1040または1010）に直接接続されることを必要とする（より詳細な例が図4を参照して後で説明される）。しかしながら一般にクライアント（1031、1032、1021、1022）はスイッチング型のネットワークを介して直接に接続される必要はなく、これらは任意の技術を用いてインタネットに接続されても良い。

【0021】2. 制御エンジンと同じ場所：クラスタとなったND-FEおよびND-CEの組は従来型のスイッチ組織または企業スイッチング型のネットワークを介して接続されうる。このやり方はクライアントおよびWANの基幹を変更しないと言う利点を持つ。しかしながら、WAN基幹がスイッチング型のネットワーク104

0であるならば、この手法はスイッチング型WAN基幹1040を完全には利用しないことになる。それでもこの手法は転送エンジン1014の分散を可能にすると共に企業スイッチング型のネットワーク（併合されたND-FE/ND-CEとサーバとの間）の利点を依然活用する。併合されたND-FE/ND-CEの例は図3を参照して後で説明される。

【0022】好適なTCP接続ルータの実施例では、FEは「実行者」プロセスを含み、CEは「実行者」および「管理者」プロセスを含む。これらは例えばNDから適応されたものである。「実行者」は高速IPパケット転送を支援するOSカーネルの拡張であって良く、また「管理者」は「実行者」を制御するユーザ・レベルのプロセスであって良い。この新しい実施態様はその「実行者」を有するFEがスイッチまたはスイッチング型のネットワークの端に分散されることを可能にして性能および頑強性の改善をもたらす。FEはCEの指示の下で同期的または非同期的に接続を分散する。ホストは従来から通信していたのと同じ方法でCEと通信することもできる。一次と指定されるCEおよび二次と指定されるもう1つのCEと言う2つのCEをネットワークに持つことによって高い利用可能性およびフォールト・トレランスが得られる。2つのCE間の通信は公知の方法を用いることができる。FEが分散されているので相違点としては一次とバックアップとの間で追加の状態が転送されねばならないことが挙げられる。

【0023】従来型のネットワーク・ディスパッチャまたはルータとは異なり、本発明はトポロジ、速度、リンク速度、およびクライアントの要求をサーバに伝えるためにスイッチング型のネットワークに備えられるその他の情報を用いることができる。例えば、ATMなどのスイッチング型のネットワークではPNNI（Private Network to Network Interface）などの経路指定プロトコルがスイッチに大量の情報を分配し、これがND-CEやND-FEにより用いられて適当なサーバに至る適当なルートを選択する。このようなルート決定プロトコルを利用する場合、ND-CEおよびND-FEは、サーバ選択を改善するのに用いられるその他のスイッチ、それらの間のリンク、および装置に関するおそらくは多様な関連情報にアクセスする。このような有用な情報の例には、端から端までの遅延（サーバに至るまで）ジッター（サーバの至るまでの遅延の変動）、サーバへの、およびサーバからのスループット（平均、ピーク、バースト度）、および伝播遅延が含まれる。これは構成可能な計測量が最適なサーバを選択するために基準となることを可能にする。この決定はND-CEにおいてなされ、ND-CEは、重みおよびND-FEがその基準を用いることができる条件などの決定基準をND-FE1013、1014に送る。このようにしてND-FEはそれが作った現存するスイッチ接続を用いてサーバに接続要

求を独立的に割り当てることができる。ND-CEのクライアントであるND-FEの部分がこの情報を優先的に受け取ってND-FEにおけるその使用を指示することになることは当業者にとって明らかであろう。

【0024】図1に戻ると、1つ以上のクライアント1021、1022がSWAN (Switched Wide Area Network) 基幹1040にルートされた基幹1020 (ルートされたネットワークとも呼ばれる) およびES/ND-FE (Edge Switch/Network Dispatcher Forwarding Engine) 1014を介して接続されている。クライアント1031、1032はルートされたネットワーク1030およびES/ND-FE 1013を介してSWAN基幹1040に接続されている。ルートされた基幹1020、1030は、クライアントの要求を端スイッチ1013または1014に送ることのできる任意の基幹として広義に定義される。スイッチされる企業ネットワーク1010もスイッチ1012を介してSWAN基幹1040に接続される。ネットワーク・ディスパッチャ制御エンジンND-CE 1011もSWAN基幹1040に接続される。サーバのクラスタ1005はスイッチされる企業ネットワーク1010に接続される。

【0025】ネットワーク1040および1010がATMネットワークである場合にはNHRP (Next Hop Resolution Protocol) が本発明に従って使用される。NHRPは同じサブネットに属さない端点間にいわゆるショートカット接続を作り、中間のすべてのNBMA (non-broadcast multiple access) を取り付けられたルータをバイパスすることを可能にする。標準的なNHRP要素はNHRPクライアント (NHC) およびNHRPサーバ (NHS) を含む。この場合、ND-FEおよびND-CEはNHC++ (以下に説明する追加の機能を有する標準的なNHC) と呼ばれる変更されたNHRPを用いることができる。一実施例において、ND-CE 1011はNHS++ (以下に説明する追加の機能を有する標準的なNHS) と呼ばれる変更されたNHRPを含んでいる。

【0026】サーバのクラスタ1005には幾つかの目標となりうるものがあるが、クライアントはただ1つの目標IPアドレス (仮想的なカプセル化されたクラスタ (VEC) アドレス) しか見ていないので、NHRPの具現化には特別の拡張が加えられなければならない。従ってNHC++およびNHS++機能 (以下に述べる) がこの場合設けられなければならない。必要とされる特定の装置の数は顧客のどのような構成でも支援できるように最小にされるべきである。ショートカット接続の利点を最大限に利用するために、好適な実施例はNHC++クライアントをSWAN 1040への入り口ルータに置き、NHS++機能を少なくともND-CE 1011に置く。WAN入り口NHC++からND-CE/NHS++への経路にあるルータに必要とされるすべてのこ

とはNHSの支援である。これと同様に、ND-CE/NHS++から目標サーバのNHC++への経路にあるルータに必要とされるすべてのことはNHSの支援である (図7を参照してこの後述べられるように)。企業ルータ (WAN出口1) はNHS++機能を持つ必要はない。

【0027】図2は本発明に従ってクライアントとサーバとの間でスイッチング経路を設定するための論理フローの例を示す。ここに示されたように、クライアント1021はサーバのクラスタ1005からTCPサービスを得るための要求2035を出す。この要求2035は、スイッチング型ワイド・エリア・ネットワーク基幹1040の境界にある端スイッチ1014に最終的に到達する。この端スイッチ1014はネットワーク・ディスパッチャ転送エンジン (ND-FE) 2023も含んでいる。ND-FE 2023は、要求が現存する接続の一部であるか否かを調べるために標準的なテーブル・ルックアップを行う。要求が現存する接続の一部であるならば、ND-FE 2023は対応するスイッチング型接続を取り出して、現存するスイッチング型接続上のクライアント要求をクラスタ1005のサーバにそのまま送る (スイッチング型のネットワーク1040、スイッチ1012、およびスイッチング型企業ネットワーク1010を介して)。既に存在する接続がないならば、ND-FE 2023はクライアント要求2025をネットワーク・ディスパッチャ制御エンジンND-CE 1011に送る。ND-CE 1011はクラスタ1005においてサーバを選択し、選択されたサーバのスイッチング・アドレスをND-FE 2023に戻す。

【0028】これに加えて、ND-CE 1011はその決定基準 (重みなどの) およびND-FE 2023がこの基準を用いることのできる条件をND-FE 2023に送ることができる。ND-FE 2023は、それが作った現存するスイッチ接続を用いてサーバ1005に接続要求を独立して割り当てるためにこの基準を用いることができる。また、ND-CE 1011は、サーバへのスイッチング経路をそれがアイドルになった後どのくらい長く維持するかについての情報をND-FE 2023に与えることが望ましい。この追加の情報は同じフロー (2026) を用いて、または別途 (2029) 送ることができる。

【0029】接続設定に戻ると、ND-CE 1011は、スイッチング型のネットワーク1040、スイッチ1012、およびスイッチング型企業ネットワーク1010を介してクラスタのサーバ1005に最初のクライアント要求を送る (2015)。ND-FE 2023が選ばれたサーバのスイッチング・アドレスを受け取ると、それは選択されたサーバへのスイッチング型接続2028を作ることになる (2027)。スイッチ接続が既に存在するならば (2028)、新しく接続を作る代

わりに(2027)現存する(2028)接続をそれが再使用することが望ましい。接続が作られた後(新たな接続かまたは現存する接続)、ND-FE2023は、スイッチング型のネットワーク1040、スイッチ1012、およびスイッチング型のネットワーク1010を介して、作られたスイッチ接続2028上のクライアント接続のすべての後続パケット2036をサーバ1005に送ることになる。

【0030】クライアント1021がサーバ1005への接続を終了すると、ND-FE2023はその接続を除去としてマークし、接続終了パケットおよびその接続に対するすべての後続パケットをND-CE1011に送るか(2025)、または接続終了パケットをサーバ1005に送り、接続が終了した後、ND-CE1011に接続が終了したことを別途通知する(2029)。接続終了パケットおよび後続パケットがND-CE1011に送られようとするとき(2025)、ND-CE1011はこの接続を除去としてマークし、パケットを関連するサーバに送る(2015)。接続が或る時間の間アイドルになっているとND-CE1011はこの接続を除去するが、これは構成可能であることが望ましい。ND-FE2023がND-CE1011に接続終了を別途通知すると(2029)、ND-CE1011はその接続テーブルから単純にその接続を除去するだけとなる。接続が終了すると、端スイッチ1014とサーバ1005との間のスイッチング型接続は、同じND-FE2023から同じサーバに向けられる追加の接続がそれを再使用することができるように維持される。どのクライアント接続もそのサーバを要求することなく或る時間が経過した後、対応するスイッチ接続は除去されても良い。

【0031】図3は、ND-FE、ND-CE、端スイッチ、企業基幹へのスイッチが同じ物理的ボックスまたは装置に取り込まれた例を示す。図に示すように、クライアント3080、3081、3082はルート指定されたネットワーク3110、ND-CE-FEスイッチ3010、およびスイッチング型のネットワーク3100を介してサーバ3090、3091、3092のクラスタにアクセスすることができる。クライアント3080からの最初の要求3005はND-FE3020に到達する。ND-FE3020はその接続テーブルでのルックアップが不首尾となった後、最初のクライアント要求3005をND-CE3040に送る(3025)。ND-CE3040はクラスタからの選択されたサーバ3090のスイッチング・アドレスで以てND-FE3020に回答する。ND-CE3040はまた最初のクライアント要求3005を選択されたサーバ3090にも送る(3035)。同じクライアント接続上でクライアント3080により出されるすべての後続パケット3045はクライアント3080からND-FE3020

にルートされ、そしてND-FE3020から選択されたサーバ3090にスイッチされる(3055)。終了のフローは図2に述べたのと同じである。

【0032】図4はND-FE1014および図1のクライアント1021が併合されてクライアント/ND-FE4420となった例を示す。図に示されたように、クライアント/ND-FE4420が新たな要求を出すと、それはSWAN1040を介して直接(4420)ND-CE1011に行き、ND-CE1011はクラスタの選択されたサーバ1005のスイッチング・アドレスをクライアント/ND-FE4420に戻す(4415)。ND-CE1011は選択されたサーバ1005にも要求を送る(4425)。この最初の交換の後、この接続についてのすべてのトラフィックはスイッチされる基幹1040を介してクライアント/ND-FE4420と選択されたサーバ1005との間でスイッチされる(4445)。終了のフローは図2について述べたものと同じである。

【0033】図5はサーバを選択し、そのサーバへの経路を選択または設定するためにND-FEで用いられる論理フローの例を示す。ステップ5010において、ND-FE2023(図2)はクライアント要求を受け取る。ステップ5030において、ND-FE2023は、この要求が属する現存する接続があるか否かを調べるためにテーブル・ルックアップを行う。現存する接続があるならば、ステップ5170においてそれは関連のスイッチング型接続を介してこの要求を単に送るだけである。ステップ5030において現存する接続がなければ、ステップ5060においてND-FE2023は、宛先サーバをローカルで選択することができるか否か、またはND-CE1011まで行かなければならないかどうかを調べる(図2)。この決定は前のフロー(2029-図2で述べた)でND-CE1011により与えられた構成可能な機能およびデータを用いてなされるのが望ましい。ND-FE2023がサーバをローカルで選択することができる場合には、ステップ5100においてそれはサーバを選択する。ステップ5060においてND-FE2023がサーバをローカルで選択することができない場合には、ステップ5090においてND-FE2023はND-CE1011を調べてサーバ選択および対応するスイッチング・アドレスを得る(図2の2025)。サーバがローカルで、またはND-CE1011により選択されると、処理はステップ5130に続く。ステップ5130において、ND-FE2023は、選択されたサーバへの現存するスイッチング型接続が存在するか否かを調べる。スイッチング型接続があるならば、ステップ5170においてそれは現存するスイッチング型接続を通してこの要求を送る(図2の2028)。サーバへの現存するスイッチング型接続が存在しないならば、ステップ5160においてND-FEは

選択されたサーバへのスイッチング型接続を設定する。これがなされると、ステップ5170においてそれは新たなスイッチされる(図2の2027)接続を通してクライアント要求を送る。

【0034】図6ないし12は、割り当てられたサーバをND-FEに知らせてそれがスイッチされる(ATM)接続にTCP接続をマップすることができるようにするためND-CEとND-FEとの間でNHRPプロトコルを用いる本発明の例を示す。本発明の一部はND-FE、ND-CE、およびサーバの間の経路に関し、またサーバ上でNHRPの公知の特徴を用いる。本発明がその他のタイプのスイッチ、またはスイッチング型のネットワークにも直ちに実施できることは当業者にとって明らかであろう。好適な実施例に用いられるすべてのフローは標準的なNHRPフローである。NHRPは拡張フィールドを許容するが、これは追加の機能を取り入れるために用いるのに好都合である。従って、フローに関するすべてのエラーはこの分野で知られた技術を用いて具合良く処理される。

【0035】図6は本発明を取り入れたネットワーク・トポロジの例を示す。TCPクライアント(101)は、サーバ141、および142を含むサーバのクラスタのサービスを用いることを必要とするIPホストである。クライアントはサーバの1つ(141または142)上のアプリケーションとTCP接続を作らなければならない。この例において、クラスタのIPアドレスはIP_SCであるものと仮定される。TCPクライアントはクラスタのIPアドレスIP_SCおよびTCPポート番号を知っているだけである。サーバはNBMA(Non-Broadcast Multiple Access)ネットワーク(162)内に置かれる。この図および以降の図において、NBMAネットワークにおけるスイッチング型接続はNBMA接続と呼ばれることになる。ここで判るように、ネットワーク162はATMネットワークであるが、本発明がその他のタイプのスイッチング型のネットワークにおいても実施できることは当業者にとって明らかであろう。

【0036】普通は、クライアント101によってサーバの1つ141または142に送られるIPデータグラムは点線で示された経路に通常は従う。ルート指定されたこの経路は幾つかの中間のルータ、つまりサーバ141については111、131、132、121および134、サーバ142については111、131、132、121および135と交差する。

【0037】本発明によれば、ネットワーク162にわたってTCP接続の間ショートカットATM接続が作られて中間のホップ数が最小になるようにされる。

【0038】中間のルータの内の2つ111および121が本発明に従って用いられている。ルータ111はND-FEを含んでいる。それはサーバとのショートカッ

ト接続を作る。それはまた、データグラムをTCPクライアントからショートカット接続に送る。ルータ121はND-CEを含んでいる。それは各新たなTCP接続毎にサーバを割り当てる。

【0039】以下のフローはIETF(Internet Engineering Task Force)によって標準化されようとしているNHRPからのものである。NHRP要素はNHRPクライアント(NHC)およびNHRPサーバ(NHS)を含むが、これらは共にこの分野では知られている。ND-FE(111)およびND-CE(121)はNHC++(以下に述べる追加的機能を有する標準的なNHC)と呼ばれる変更されたNHRPクライアントを用いる。ルータ133もNHS++(以下に述べる追加的機能を有する標準的なNHS)と呼ばれる変更されたNHRPサーバを含んでいる。必要とされる特定の装置の数はどのような構成をも支援できるように最小でなければならない。WAN入り口NHC++111からNHS++133間での経路にあるルータ131、132に必要とされるすべてのことはNHSの支援である。

【0040】ショートカットの利点を十分に活用するための好適な実施例はNHC++クライアントを、少なくともND-CE121においてWANおよびNHS++機能に至る入り口ルータ111に置くことである(ルータ133の代わりに)。この場合、ルータ133はNHSも支援する。

【0041】これと同様に、ND-CEにサービスを与えるNHS++(133)から目標サーバ141、142への経路にあるすべてのルータ(図6には何も示されていないが、ルータ134とサーバ141の間には1つ以上のルータが存在しうる)に必要とされるすべてのことはNHSの支援である(図7に関して説明する)。企業ルータ111(WAN出口1)はNHS++機能を持つ必要がないことに留意されたい。WAN入り口および出口ルータは同じルータである必要はないことは当業者にとって明らかであろう。

【0042】図7は図6に示したネットワーク・トポロジに関する初期化フローの例を示す。

【0043】フロー201:各NHC(111, 121, 141)はそれぞれが所有するNHS(131, 133, 134)に対するNBMA接続設定を開始する。この例ではNHS131はNHC111に対するサーバであり、NHS133はNHC++121に対するサーバであり、NHS134はNHC141に対するサーバであるものと仮定される。NHCおよびそれにサービスを与えるNHSの総体的所在場所は本発明の原理に関係がない。

【0044】フロー202:フロー201が完了した後、各NHC(111, 121, 141)はそれ自身のプロトコル・アドレスおよびそれ自身のハードウェア・アドレスをそれぞれにサービスするNHS(131, 1

33、134)に登録する(NHRP REGISTER)。例えば、ND-CE121はサーバ・クラスタのIPアドレスをそれ自身のATMアドレスと共に登録する。

【0045】フロー203：アドレス登録が完了すると、NHS131、133、134は肯定応答をそのクライアントに送る。

【0046】フロー214：ND-CE121はサーバ・クラスタの各ホストとのNBMA接続を用いて受け取った最初の packets をルートを決められた経路に送る。各ホスト毎にそれは権限ある解決要求(NHRP RESOL)をそれにサービスするNHS133に送る。この解決要求は宛先ホストのIPアドレス(例えば、サーバ141に対しIP1)を指定する。この例では初期化フロー214ないし228がクラスタの1サーバについて説明される。これらのフローは各サーバ毎に実行されなければならない。

【0047】フロー215：NHS133は解決要求をその隣のNHSに送る。この分野で知られた手法を用いて要求は、要求されたIPアドレスを有するNHS、例えばIP1に対してはNHS134、に到達する。

【0048】フロー216：NHS134は141のハードウェア・アドレスを含む解決応答を要求を出した所に送る。

【0049】フロー217：この分野で知られた手法を用いて解決応答は要求を出したところ、即ちND-CE(121)、に到達する。

【0050】フロー228：ND-CE(121)はサーバ141とのATMショートカット接続をここで作り上げる。

【0051】図8は、サーバ141および142を含むクラスタとのTCP接続を作るためTCPクライアント101(図6)によって送られるIPデータグラムの処理の例を示す。この例はND-FEがローカルにサーバを選択しようとしていないときの論理フローを述べる。上に述べたように、ND-FEがサーバをローカルで選択するときには、それはNBMA接続を再使用するものでこれについてのフローは不要である。

【0052】フロー301：TCPクライアント101はIPデータグラムを送って新たなTCP接続(TCPオープン接続)を要求する。データグラムの宛先IPアドレスはIP_SC、即ちサーバ・クラスタのアドレスである。ソースIPアドレスはTCPクライアントのアドレス(IP_CL)である。TCPヘッダはソースTCPポート番号(P1)および宛先TCPポート番号(P2)を含んでいる。4つの組合せ(IP_SC, IP_CL, P2, P1)(以下、“TCP接続キー”と呼ぶ)はTCP接続を一意的に指定する。

【0053】フロー302：IPデータグラムはND-FE(111)に到達する。ND-FE(111)はそ

の「ND-FEキャッシュ・テーブル」を調べてTCP接続キー(IP_SC, IP_CL, P2, P1)に一致する項目を探す。これは新たな接続なので、このテーブルにこのような項目は無い。ND-FE(111)はこのデータグラムをデフォルト・ルートの経路303(図6の点線で表される)に送る。ND-FE(111)はまたND-FEキャッシュ・テーブルに新たなTCP接続キーに対する新たな項目を作る。このTCP接続にはNBMA接続は関連されない。

【0054】フロー303：IPデータグラムはルートされた経路(131、132および133)に沿ってすべてのルータにより送られる。

【0055】フロー304：ND-CE(121)はIPデータグラムを受け取り、そのキャッシュ・テーブルを調べてTCP接続キー(IP_SC, IP_CL, P2, P1)に一致する項目を探す。これは新たな接続なので、このテーブルにこのような項目は無い。要求されたサービス(TCP宛先ポートP2によって指示される)、およびサーバの負荷またはその他の情報に基づいて、ND-CE(121)は新たなTCP接続のためのクラスタ中の最良のサーバを決定する。選択されたサーバはサーバ141であるものとここでは仮定される。従って、ND-CE(121)は以前に作られたNBMA接続を通してIPデータグラムをサーバ141に送る(フロー228)。新たなTCP接続がサーバに対して作られると、ND-CE(121)はそれ自身のTCP接続キャッシュ・テーブルに新たな項目を加えて不活動タイマをスタートさせる。

【0056】フロー315：サーバ141との新たなTCP接続が作られようとしているので、ND-CE121は修正されたNHRP REGISTER要求を用いてこの新たなTCP接続をそれにサービスするNHS133に登録する。この修正された要求はND-CEに特有の拡張フィールドを持っており、これはTCP接続キー(IP_SC, IP_CL, P2, P1)、および宛先サーバのハードウェア・アドレス、即ちサーバ141のATMアドレスを指定する。

【0057】フロー316：アドレス登録が完了すると、NHS133は肯定応答をそのクライアントに送る。

【0058】フロー327：短い遅延の後、ND-FE(111)は修正された権限NHRP RESOLUTION要求をそれにサービスするNHS(131)に送る。この修正された要求は、TCP接続キー(IP_SC, IP_CL, P2, P1)を指定するネットワーク・ディスパッチャ特有の拡張フィールドを含むことが望ましい。ND-FEはルート指定された経路を用いてこの要求が満足されるまでクラスタにパケットを送り続けることになる(フロー303)。この要求に対して否定応答があると、ND-FE111は再び要求する。

【0059】フロー328：この要求は権限あるものである。この分野で知られた手法を用いてこの要求はNBMAネットワーク162（図6）を通して送られ、要求されたTCP接続キーを有するNHS133に到達する。

【0060】フロー329：NHS133は修正されたNHRPキャッシュにおいてキー（IP_SC, IP_CL, P2, P1）を用いてTCP接続キーを探索する。NHS133が項目を見つけると、それは選択されたサーバ141のATMアドレスを指定するNHRP RESOLUTION肯定応答を送り返す。NHS133が項目を見つけないときには、それはある時間の間遅延してNHC++/ND-CE121がNHRP REGISTERを送ることを許容する（フロー315）。遅延が終了する前にNHRP REGISTERが受け取られると、NHRP RESOLUTION肯定応答が送られ、そうでない場合にはNHS++133がNHRP RESOLUTION否定応答を要求に対して送る。

【0061】フロー330：この分野で知られた手法を用いて、NHRP RESOLUTION応答は要求者であるND-FE111に到達する。

【0062】フロー341：ND-FE（111）はTCP接続キー（IP_SC, IP_CL, P2, P1）と共にサーバ141のATMアドレスをND-FEキャッシュ・テーブルに保存し、サーバ141へのショートカットNBMA接続を作る。ATM接続ができると、それはそのインタフェース番号および標準的なATMのVPI/VC（Virtual Path Indicator/Virtual Channel Indicator）値をND-FEキャッシュ・テーブルに保存する。NBMA接続が作り上げられる前にTCP接続に対する第2のIPデータグラムがND-FE111によって受け取られると、このデータグラムはルート決定された経路に送り出される。

【0063】図9はND-FE111によるショートカットNBMA接続の使用の例を示す。

【0064】フロー401：TCPクライアント101は以前に作られたTCP接続（IP_SC, IP_CL, P2, P1）上のサーバ・クラスタにIPデータグラムを送る。

【0065】フロー402：ND-FE111はIPデータグラムを受け取る。ND-FE111はそのND-FEキャッシュ・テーブルを調べてTCP接続キー（IP_SC, IP_CL, P2, P1）に一致する項目を探索し、そしてショートカットATM接続が既に存在することを見いだす。ND-FE111はIPデータグラムをサーバ141に直接送る。

【0066】フロー410：ND-FE111はルート決定された経路を通してND-CE121のために意図されたリフレッシュ・メッセージを定期的に送る。この

メッセージは活動中のTCP接続キーのリストを含んでいる。リフレッシュ・メッセージは肯定応答を必要としない無接続データグラムであることが望ましい（例えばUDP（User Datagram Protocol）を用いて）。リフレッシュの期間はルートされるトラフィックを過剰に増大させないように十分大きく選ぶことができる。事実、ND-CE不活動タイマの期間の1/3の値で十分である。

【0067】フロー411：リフレッシュ・メッセージはサーバ・クラスタ・アドレスと同じ宛先IPアドレスを持っている。それはこの分野で知られた手法を用いてND-CE121に到達する。

【0068】図10はTCP接続が閉じられようとしていることを示すTCPパケットの処理の例を示す。

【0069】フロー501：TCPクライアント101はTCP接続終了を示すTCPパケットを送る。ND-FE111はこのパケットを受け取り、TCP接続キー（IP_SC, IP_CL, P2, P1）を取り出してND-FEキャッシュ・テーブルを調べ、TCP接続キーに一致する項目を探索する。それは接続状態を“クローズ”とマークし、ルート決定された経路にこのパケットを送り出し、そしてタイマを始動させる。このタイマは、最後のパケットが接続を通った後に、項目を追いつ前にどのくらいの長さだけ待つのかを指定する。この接続に対する以後のパケットはすべてこのルート指定された経路に送り出されるのでND-CE121は同じタイマを維持することができる。正しい動作のためには、このタイマはMSL（Maximum Segment Lifetime）の2倍よりも大きくなければならない。指示された長さに時間の間接続がアイドルになっていると、ND-FE111は対応する項目をそのND-FEキャッシュ・テーブルから取り除く。

【0070】フロー502：TCPパケット宛先アドレスはサーバ・クラスタ・アドレスであるIP_SCである。従って、このパケットはそれがND-CE121に受け取られるまでルータからルータへとルートされる。ND-CE121はTCP接続キーを取り出して対応するサーバ141のアドレスを得る。

【0071】フロー503：ND-CE121は以前に作られたATM接続（フロー228）を通してTCPパケットをサーバ141に送る。それはTCP接続の状態を“クローズ”とマークしてタイマを始動させる。このタイマもMSL TCPタイマの2倍より長い。

【0072】フロー510：タイマがタイムアウトすると、ND-CE121はそのテーブルから対応する項目を取り除く。それはまた修正されたNHRP PURGE要求（例えばTCP接続キーを含む）をそれにサービスするNHSに送る。

【0073】フロー511：NHS133はその内部テーブルからTCP接続キーを取り除く。それはまたNH

RP PURGE 応答を要求元 121 に送ることによって応答する。

【0074】図11はサーバによるショートカットATM接続をクリアする処理の例を示す。

【0075】フロー601：ショートカットATM接続はサーバ141によってクリアされる。サーバは2つの理由により自由意志でショートカット接続をクリアすることができる。タイマの或るものはタイムアウトしているかまたはそのATMアドレスが変化しようとしている。この要求はNBMA接続に関連する会話の上で流れるすべてのパケットに、NBMA接続が選択されたサーバに対して再び作られるまで、ルート決定された経路を辿らせる。

【0076】フロー610：1つ以上のTCP接続項目が、クリアされようとしているNBMA接続に関連するND-FE111 (ND-FEキャッシュ・テーブル) に存在するならば、NBMA接続が再び作られることが必要である。サーバのATMアドレスは認証される必要がある。修正された権限NHRP RESOLUTION (図8のフロー327で定義されたもの) 要求がNHS131に送られる。サーバ141およびND-CE121が再初期化するための時間を与えるために、NHRP RESOLUTION 応答はATM接続がクリアされた直後には送られないのが望ましく、それは短い遅延の後に送られる。

【0077】フロー611：この要求は権限あるものである。NHS131はこの解決要求をその隣のNHS132に送る。この要求はこの分野で知られた手法を用いて、要求されたTCP接続キーを有するNHS133に到達する。

【0078】フロー612：NHS133はその修正されたNHRPキャッシュにTCP接続キー (IP_SC, IP_CL, P2, P1) を見つけている。NHS133は、選択されたサーバ141のATMアドレスを指定するNHRP RESOLUTION 肯定応答を送り返す。

【0079】フロー613：NHRP RESOLUTION 応答はこの分野で知られた手法を用いて要求元のND-FE111に到達する。

【0080】フロー620：ND-FE111はサーバ141に対するATMショートカット接続を再び作る。

【0081】図12はネットワークによりATMショートカット接続をクリアする処理の例を示す。

【0082】フロー701：ND-FE111とサーバ141との間のショートカットATM接続はネットワークによりクリアされる。ND-FE111はこの接続をND-FEキャッシュ・テーブル中のどの項目によっても使用できないものとしてマークする。これは、この接続を用いたであろう後続のすべてのパケットを、選択されたサーバに対してNBMA接続が再び作られるまで、

ルート決定された経路を通して送らせる。

【0083】フロー710：ND-FE111はサーバ141とのショートカット接続を再び作ろうと試みる。

【0084】高度の利用可能度およびフォールト・トレランス

図1は本発明の高度の利用可能度の特徴をも示す。スイッチング型のネットワーク1040の内部には1つ以上のND-CEが存在しうる。これらのCEはそれらの内部テーブルを同期状態に保つために上述のものと同一キャッシュに一貫性あるプロトコルを用いることになる。ND-CEが障害を起こすと、それを引き継いだND-CEが、新たな制御ND-CEであることをすべてのND-FEに知らせる。

【0085】ND-FEはND-CEとは無関係に障害を起こす。ND-FEが障害を起こすと、この障害を起こしたND-FEを介して接続されたクライアント (1021, 1022, 1031または1032) だけが影響を受ける。図1において、クライアント1021, 1022はND-FE1014を介して接続され、クライアント1032, 1031はND-FE1013を介して接続されている。ND-FE1013が障害を起こすと、期間1030を介して接続されたクライアント (1032, 1031) だけが影響を受ける。クライアントの要求を本発明のシステムに導く期間ネットワーク (ルートを与える期間) が1つのND-FE1013だけを有しそれを介してネットワークがルートを与える場合には、そのFEの障害はこれらのクライアントを恒久的に切り離すことになる。この1点の障害から保護するために、第2のND-FEがこのルート基幹1030に取り付けられるように構成に組み込まれることができる。しかしながら、典型的には、インターネットのようなルート基幹では複数のルートが利用可能となる。クライアントが接続すると、ネットワークにおいて得られるルート情報が、ND-FE1013が障害を起こしたときに別のどのND-FE (またはルート) が選択されうるのかを決めるのに使用されうる。一次および二次ND-FE優先的選択情報を用いてこのように構成されうる。接続テーブルを活動中のND-FEと同期するよう維持するために一次ND-FEと次の2つの最も可能性あるND-FEとの間にキャッシュ一貫性プロトコルが維持されうる。キャッシュ一貫性プロトコルはTCP接続キーおよび選択されたサーバの識別子をバックアップのND-CEに送る。バックアップのND-CEがTCP接続キーを受け取ると、ショートカットが割り当てられ、その接続テーブルに項目が記入される。バックアップのND-FEと選択されたサーバとの間にショートカットが存在しない場合、ショートカットが作られる。最初のパケットが受け取られるまでこの接続を作るのを遅らせるために既に述べた定義のフローを用いることができることは当業者にとって明らかであろう。図1において、ND-

FE1013を一次ND-FEとし、ND-FE1014をバックアップとして考察する。ND-FE1014接続テーブルはND-FE1013(1014はこれの代替として構成に入れられている)を介して接続されたすべてのクライアントについての項目を含んでいる。一次ND-FEが障害を起こすと、基幹はパケットを構成に入れられた代替のND-FEにルート指定する。ここで、代替を構成に組み入れることは、ルーティング・テーブルなどのネットワーク構成に関する従来の手法を意味する。しかしながら、その他の構成機構も利用可能であることは当業者にとって明らかであろう。例えば、ネットワーク・トポロジを動的に決定する方法が知られている。この知識は一次およびバックアップを構成するのに用いることができる。

【0086】再び図5を参照すると、転送決定論理に関連した、すべての接続、一次およびバックアップは同じ接続テーブルで与えられることが望ましい。

【0087】一次ND-FEが通常行われるように修復または回復したとき、それはルートが利用可能であるかどうかについてネットワーク1030を更新する。それは、そのバックアップとして構成に入れられたFEからキャッシュ更新を得ることになる。通常行われるように、ネットワークはルートのためのND-FEの使用可能性を知るので、新たな接続が自動的にそれにルートされ、現存する接続に対するパケットはクライアントに対する中断なしにND-FEを介して再ルートされることができるようになる。一次ND-FEへのスイッチ・バックはこの分野で知られた標準的なプロトコルによって処理することができる。ND-FE間のキャッシュ貫性プロトコルは本発明がパケット・スイッチング型のネットワークの利点を利用することを可能にする。

【0088】ND-FE間のキャッシュ貫性プロトコルは活動中の接続が終了したことがバックアップであるND-FEに同報されることを保証する。輻輳状態の変動のためにネットワークがパケットを異なるやり方でルートするならば、2つのND-FEに1つのクライアントが活動しているものと映ることがある。或る点で接続

が終了され、そしてこれらND-FEの各々は他のND-FEのバックアップであるため、終了はキャッシュ貫性プロトコルを介してバックアップND-FEに繰り返されることになってバックアップFEからクライアント接続が外されることになる。

【0089】本発明は好適な実施例によりその代替実施例と共に説明されたので、本明細書の特許請求の範囲に含まれる種々の等価物、改良、改善が現在、および将来においても当業者にとって自明であることが理解されよう。従って、特許請求の範囲は最初に開示された本発明の適正な保護を全うするように解釈されなければならない。

【図面の簡単な説明】

【図1】 スwitchング型のネットワーク配置された本発明の1例を示す図。

【図2】 スwitchング経路を設定するための論理的フローの1例を示す図。

【図3】 単一のスイッチに併合された制御エンジン(CE)および転送エンジン(FE)の例を示す図。

【図4】 クライアントに併合されたFEの例を示す図

【図5】 FEにおける決定プロセスの例を示す図。

【図6】 ネクスト・ホップ・レゾリューション・プロトコル(NHRP)を用いたネットワーク・トポロジの例を示す図。

【図7】 種々の要素に対するNHRP初期化フローの例を示す図。

【図8】 TCP接続設定のためのクライアント・サーバ論理フローの例を示す図。

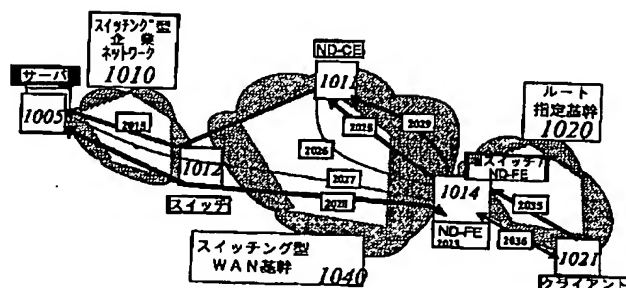
【図9】 TCP定常状態に対するクライアント・サーバ・フローの例を示す図。

【図10】 TCP閉鎖接続のためのクライアント・サーバ・フローの例を示す図。

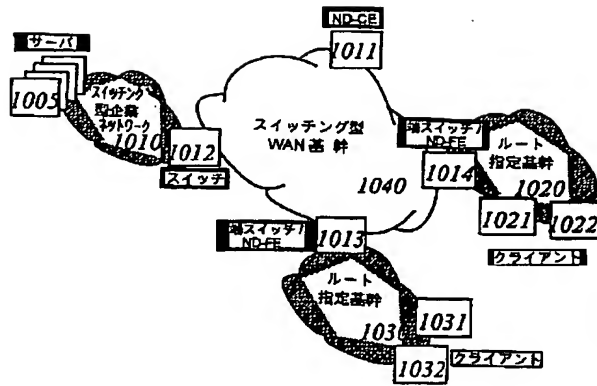
【図11】 サーバによりクリアされたクライアント・サーバ・ショートカット接続の例を示す図。

【図12】 ネットワークによりクリアされたクライアント・サーバ・ショートカット接続の例を示す図。

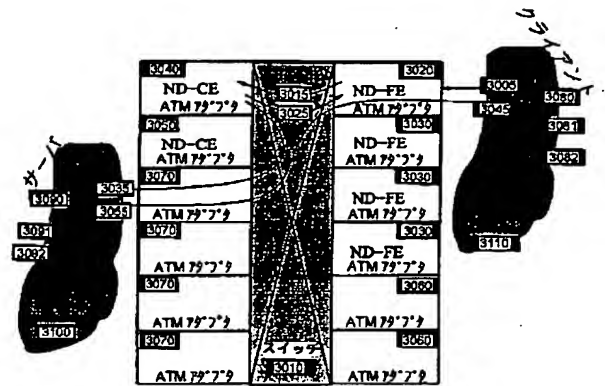
【図2】



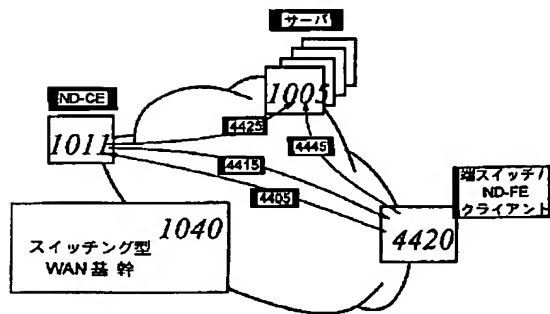
【図1】



【図3】

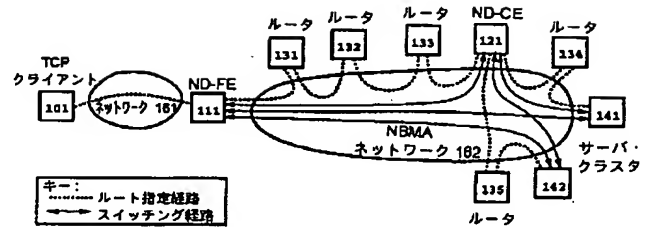


【図4】



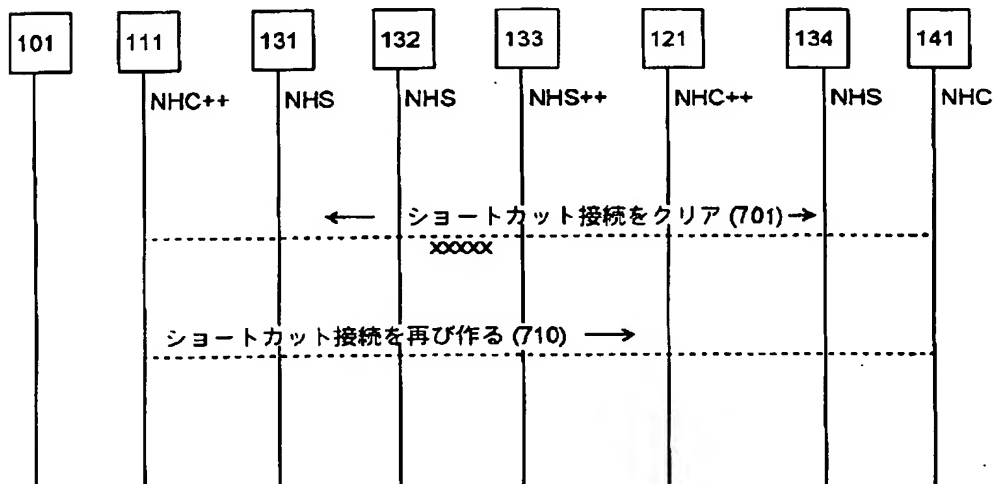
【図6】

ネットワーク・トポロジの例

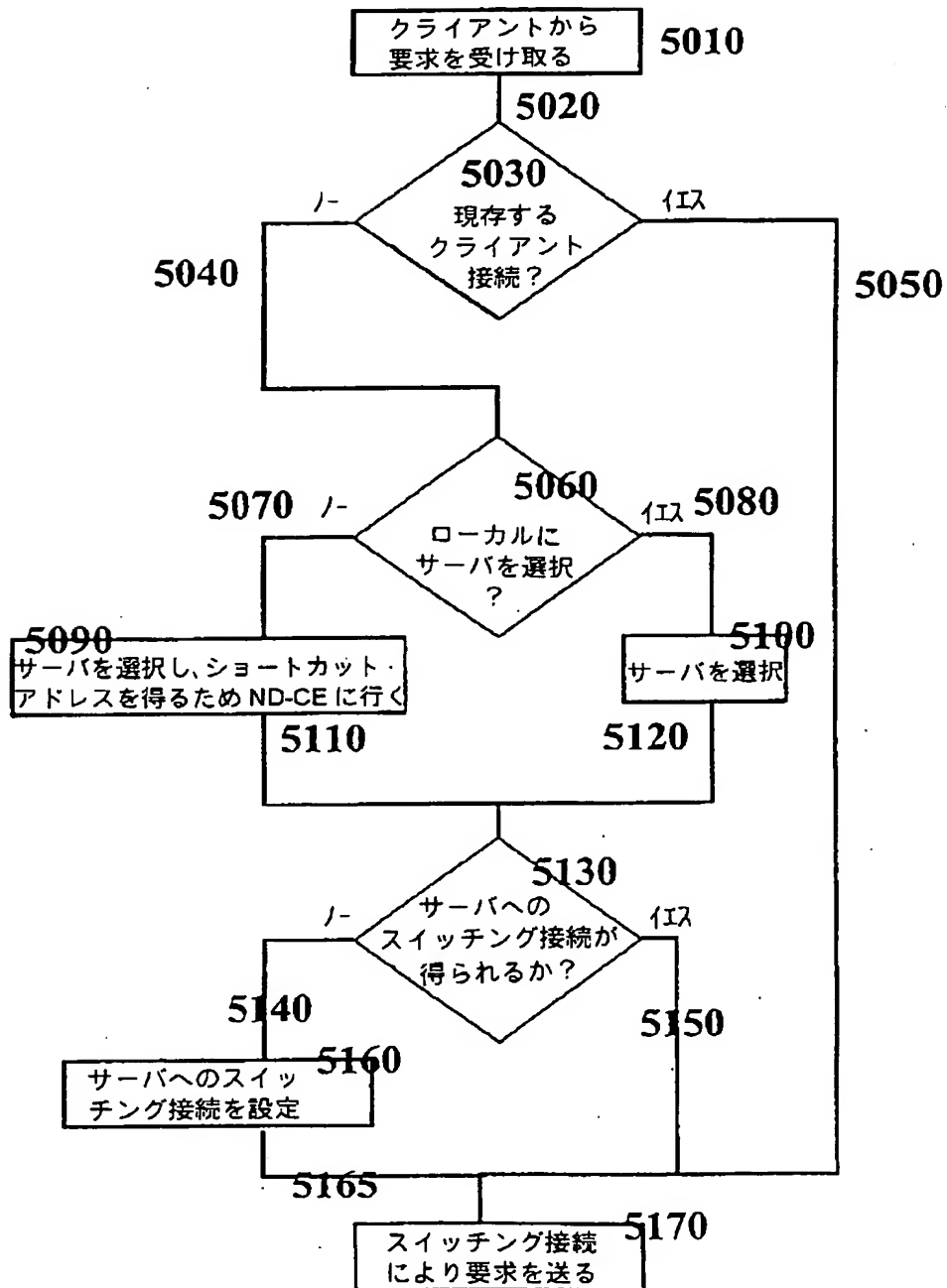


【図12】

ATM接続をネットワークによりクリア

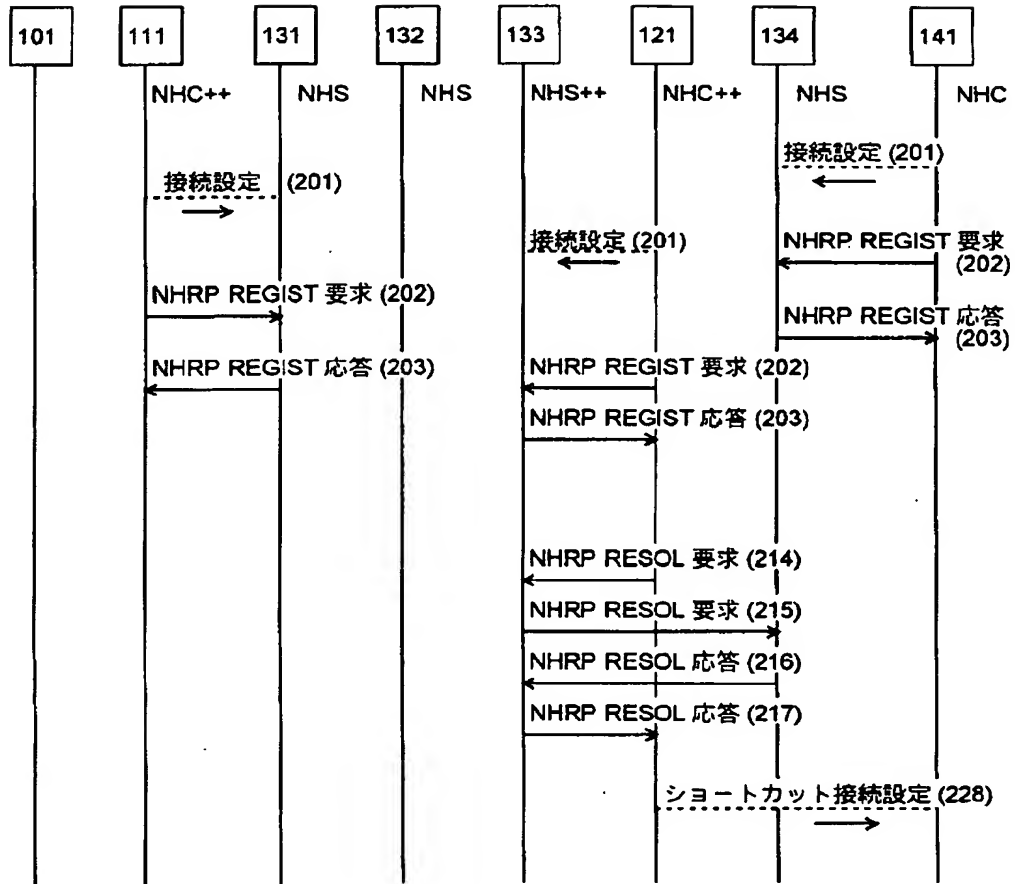


【図5】



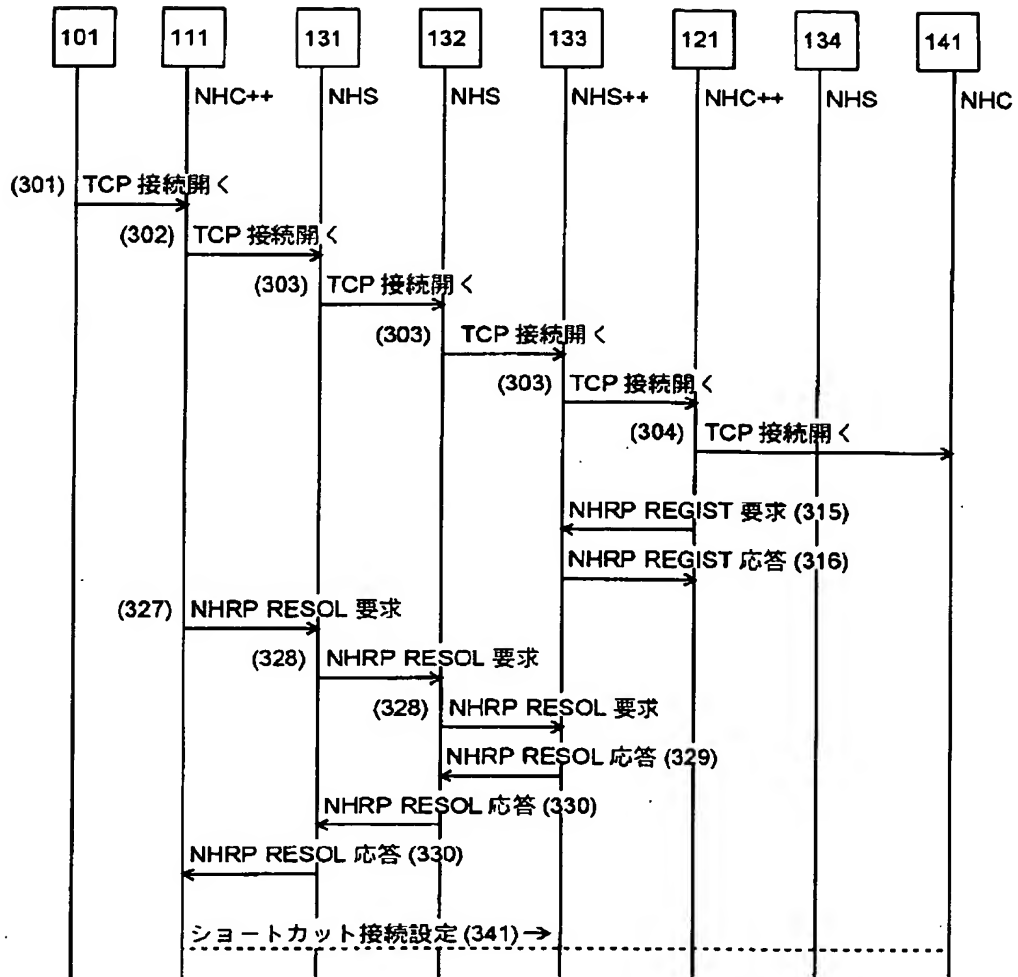
【図7】

初期化フロー



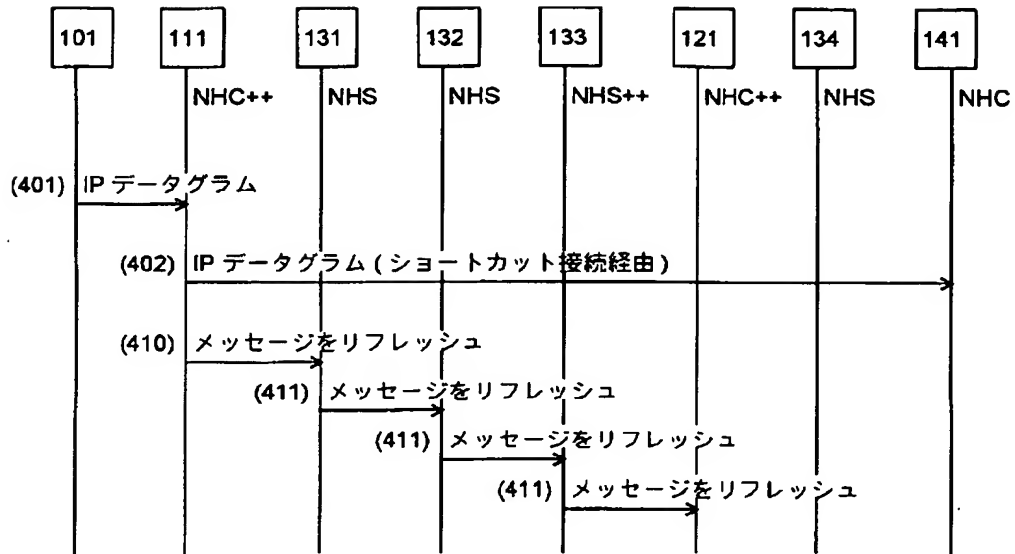
【図8】

T C P接続設定



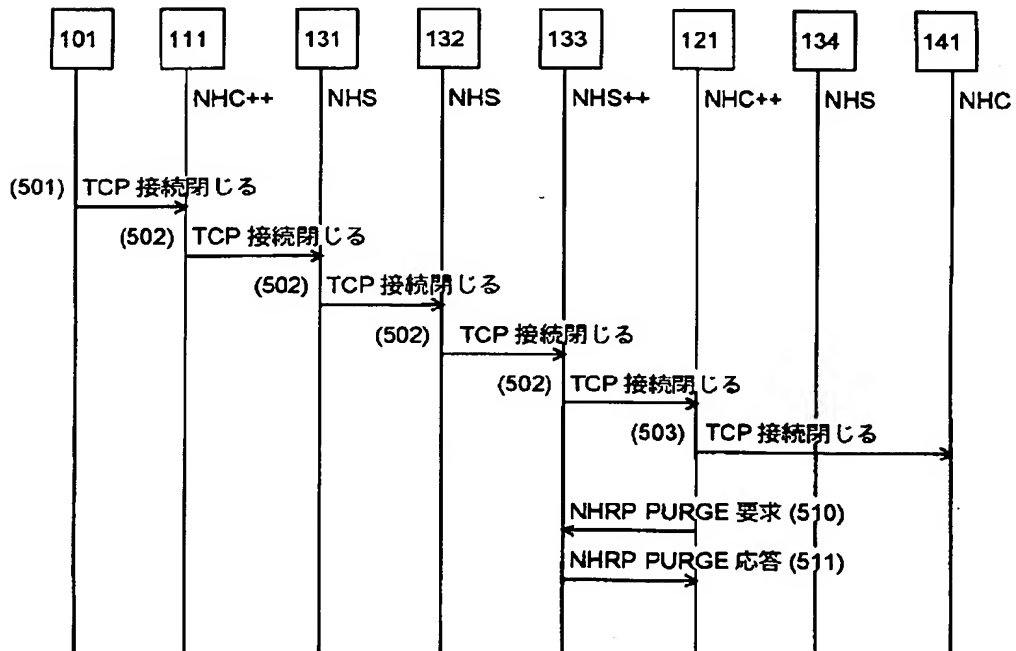
【図9】

TCP 接続一定常状態



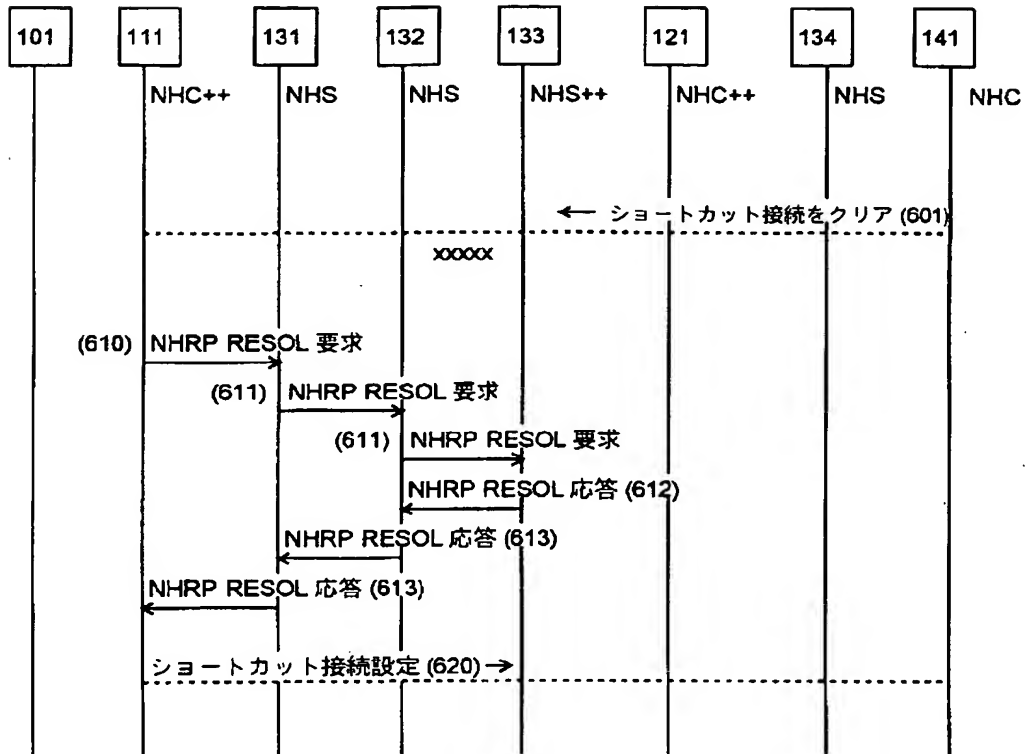
【図10】

TCP 接続を閉じる



【図11】

サーバによりATM接続をクリア



フロントページの続き

(72)発明者 ガルニー・ダグラス・ホロウェイ・ハント
アメリカ合衆国10598、ニューヨーク州ヨ
ークタウン・ハイツ、ウェリントン・コー
ト 31

(72)発明者 エリック・ミシェル・レヴィーアベノール
フランス国06200、ニース、アンシャン・
シュマン・ドゥ・ラ・ランテール 67

(72)発明者 ダニエル・ジョルジュ・ジャン・マリー・
モデュ
フランス国06200、ニース、1・シュマ
ン・ドゥ・ラ・バテリイ・リュス、ジャル
ダン・デ・イリ、エイ・5

Fターム(参考) 5K030 GA04 GA11 HA08 HC06 HC13
JT02 KA05 LB05 LB19 LE03
MB01 MD02 MD08
5K033 AA04 AA09 BA04 CB01 CB06
DA06 DB12 DB14 DB16 DB18
EA04 EB06 EC04